BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 4: e59077
doi: 10.3897/biss.4.59077

OPEN ACCESS

Conference Abstract

Historical Overview of the Development of the Symbiota Specimen Management Software and Review of the Interoperability Challenges and Opportunities Informing Future Development

Edward Gilbert‡, Nico Franz‡, Beckett Sterner‡

‡ Arizona State University, Tempe, United States of America

Corresponding author: Edward Gilbert (egbot@asu.edu)

Received: 28 Sep 2020 | Published: 30 Sep 2020

Citation: Gilbert E, Franz N, Sterner B (2020) Historical Overview of the Development of the Symbiota Specimen Management Software and Review of the Interoperability Challenges and Opportunities Informing Future Development. Biodiversity Information Science and Standards 4: e59077. https://doi.org/10.3897/biss.4.59077

Abstract

Symbiota (Gries et al. 2014) is an open-source software platform designed to function as a biodiversity Content Management System (CMS) for specimen-based datasets. Primarily in North America though also increasingly on other continents, the Symbiota software platform has risen to prominence in the past ten years as one of the more heavily accessed mid-level aggregation tools for assembling, managing, and distributing datasets associated with biological collections. There are more than 50 public Symbiota portals being managed and promoted by various biodiversity projects and communities. Together, these portals assist in the distribution and mobilization of more than 55 million specimen and 20 million image records associated with hundreds of institutions.

The central premise of a standard Symbiota installation is to function as a mini-aggregator capable of integrating multiple occurrence datasets that collectively represent a community-based research data perspective. Datasets are typically limited to geographic and taxonomic scopes that best represent the community of researchers leading the project. Symbiota portals often publish "snapshot records" that originate from external management systems but otherwise align with the portal's community of practice and data

focus. Specimen management tools integrated into the Symbiota platform also support the ability to manage occurrence data directly within the portal as “live datasets”. The software has become widely adopted as a data management platform. Approximately 550 specimen datasets consisting of more than 14 million specimen records are being directly managed within a portal instance. The appeal of Symbiota as an occurrence management tool is also exemplified by the fact that 18 of the 30 federally funded Thematic Collection Networks (https://www.idigbio.org/content/thematic-collections-networks) have elected to use Symbiota as their central data management system.

Symbiota's well-developed data ingestion tools, coupled with the ability to store import profile definitions, allows data snapshots to be partially coordinated with source data managed within a variety of remote systems such as Specify (https://specifysoftware.org), EMu (https://emu.axiell.com), Integrated Publishing Toolkit (IPT, https://gbif.org/ipt) publishers, as well as other Symbiota instances. As with Global Biodiversity Information Facility (GBIF) and Integrated Digitized Biocollections (iDigBio) publishing models, data snapshots are periodically refreshed, based on transfer protocols compliant with Darwin Core (DwC) data exchange standards. The Symbiota data management tools provide the means for the community of experts running the portal to annotate and augment snapshot datasets with the goal of improving the overall fitness-for-use of the aggregated dataset. Even though a data refresh from the source dataset would effectively replace the data improvement with the original flawed data, the system’s ability to maintain data versioning of all annotations made within the portal allows data improvements to be reapplied. However, inadequate support for bi-directional data flow between the portal and the source collection effectively isolates the annotations within the portal.

On one hand, the mini-aggregator model of Symbiota can be viewed as compounding the further fragmentation of occurrence data. Rather than conforming to the vision of pushing data from the source, to the global aggregators and ultimately the research community, specimen data are being pushed from source collections to a growing array of mini-aggregators. On the other hand, community portals have the ability to incentivize experts and enthusiasts to publish high-quality, "data-intelligent" biodiversity data products with the potential of channeling data improvements back to the source.

This presentation will begin with a historical review of the development of the Symbiota model including major shifts in the evolution of the development goals. We will discuss the benefits and shortcomings of the data model and provide a description of schema modifications that are currently in development. We will also discuss the successes and challenges associated with building data commons directly associated with communities of researchers. We will address the software’s role in mobilizing occurrence data within North America and the efficacy of adhering to the FAIR use principles of making datasets findable, accessible, interoperable, and reusable (Wilkinson et al. 2016). Finally, we will discuss interoperability developments that we hope will improve the flow of data annotations between decentralized networks of data portals and the original data providers at the source.

Keywords

biodiversity data, natural history collections, data coordination, de-centralization

Presenting author

Edward Gilbert

Presented at

TDWG 2020

References

- Gries C, Gilbert E, Franz N (2014) Symbiota – A virtual platform for creating voucher-based biodiversity information communities. Biodiversity Data Journal 2 https://doi.org/10.3897/bdj.2.e1114
- Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3: 160018. https://doi.org/10.1038/sdata.2016.18